ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

AI regulation needs statistics

Evaluating reliability, uncertainty and risk in practice

8 July 2026

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

1. Introduction

There are two competing narratives around AI systems. The first focuses on AI as a potentially transformative technology that could improve productivity, boost the UK's sluggish economic growth and support more responsive public services. The other is more cautious: AI is a technology many people do not yet fully trust, especially when it is used in high-stakes settings and governed by rules that remain partial, uneven, or unclear.[1] That tension matters. Where evidence is limited and oversight still developing, scepticism is not an irrational obstacle to innovation but a predictable response to uncertainty.

This is the context in which AI regulation must now operate. Policymakers are being asked to govern a fast-moving, probabilistic, and increasingly embedded set of technologies before robust norms of evaluation, assurance and accountability are fully in place. In the UK, the response has been to rely on existing regulators to apply broad cross-sector principles such as safety, fairness, and transparency, rather than to create a single dedicated AI regulator.[2] That approach has practical advantages, but it also raises familiar questions about consistency, capability and coverage as AI systems become more complex and more deeply woven into everyday decision-making.

This differs from the approach of the EU, which – through its AI Act – has introduced a more centralised, risk-based regulatory framework. The Act identifies certain uses of AI as high-risk and requires providers to meet specific requirements relating to risk management, transparency, documentation, human oversight and performance before systems can be put into service.

In this paper we are not intending to adjudicate between these two approaches to regulation. There is a clear sense in which the UK government's model reflects a sensible intuition: AI is not one technology, and its risks do not look the same in healthcare, education, or financial services. Our focus is on helping to make sector specific oversight work. This needs regulators who are equipped to judge whether AI systems are performing safely and effectively in practice – and understanding AI models is a very specialised skill that does not naturally exist in the sectors where they are used.

This approach to regulating AI does, though, raise questions about responsibility and accountability. AI is routinely developed by one organisation, deployed by another and then relied on by people who have limited or no sight of how its outputs are produced. Assigning responsibility for outcomes quickly becomes challenging: does it sit with the people who developed the AI, the organisation who purchased the system, or the people using it to inform decisions? This makes effective evaluation especially important – where accountability is shared in this way, evidence about the behaviour of systems is critical for regulatory oversight.

As the RSS, we argue that the statistical character of AI remains under-recognised in regulatory debate. As we set out in *AI is Statistics*, many of the questions that matter most for regulation are statistical at their core: how accurate and reliable outputs are, how performance varies across groups and settings, how sensitive systems are to shifts in data or context, how the output can be explained and interpreted and how uncertainty is measured and communicated. If AI is to be regulated well, it must be evaluated

[1] See DSIT's *Public attitudes to data and AI: Tracker survey (Wave 4)*, the Ada Lovelace Institute and Alan Turning Institute's *How do people feel about AI?*, and KPMG's *UK attitudes to AI*.
[2] See DSIT's AI regulation: a pro-innovation approach and A pro-innovation approach to AI regulation: government response.

well. And if it is to be evaluated well, statistical reasoning must sit much closer to the centre of regulatory practice.

This matters because AI systems are not static products that can be tested once and signed off. They are often adaptive and highly sensitive to the environments in which they are deployed. Performance can drift, outputs can vary across populations, and systems that appear robust in one context can fail in another. Models are now trained by a small number of resource-rich companies, away from the use cases where they are deployed. Frequent updates introduce changes in behaviour and practitioners need to be constantly vigilant. One-off approval is therefore often a weak form of oversight. What is needed instead is more continuous, statistically grounded monitoring and evaluation, an argument increasingly reflected in UK debates on AI assurance, testing and oversight.[3]

We develop this argument through three case studies: healthcare data, education, and financial services. These are very different sectors, and that is precisely the point. Each reveals a different facet of the same regulatory problem: AI is already being deployed in domains where mistakes matter, yet the methods used to evaluate and oversee these systems often lag behind the systems themselves. Taken together, the cases suggest that traditional regulatory approaches are poorly matched to technologies whose behaviour is uncertain, variable and data dependent. A more credible model of AI regulation will therefore need to embed statistical scrutiny far more deeply than most current frameworks do.

Drawing from these case studies, we make five key recommendations.

Recommendation 1: Establish clear responsibility for identifying and addressing gaps in AI regulation

As AI is adopted across sectors, it is increasingly being used in ways that cut across existing regulatory boundaries. We look at the example of the use of AI within teaching and assessment and detail how it is not always clear which body is responsible for assessing how such systems are used in practice, or for ensuring that they are sufficiently reliable and appropriate for their intended purpose. Without a lead regulator and with the AI Security Institute (AISI) focused on existential risk, there is no part of government with a specific remit to systematically identify regulatory gaps and ensure that regulatory functions are allocated and properly resourced. Some part of government – either in the Department for Science, Innovation and Technology (DSIT) or the AISI – should be given clear responsibility for identifying and addressing these gaps. They should be responsible for: 1) identifying where a use case falls outside every existing regulator's remit; 2) deciding which regulator it should sit within (or flagging that none currently fits); and, 3) checking on a defined cycle whether that regulator has the statistical capability to discharge its responsibility. Statistical expertise must also be embedded within the function for identifying regulatory gaps – since, as we show, many regulatory challenges are driven by statistical features of AI.

Recommendation 2: Ensure that the evaluation of AI systems reflects how they operate in practice

Across the case studies, we show how current approaches to evaluating AI systems often fail to reflect how they are used in real-world settings. As a result, systems may appear to perform well under test conditions while behaving unreliably in practice. Regulators should therefore set clear expectations that AI systems are evaluated in the contexts in which they are used, including how performance varies across groups, changes over time, and depends on the interactions between data, models and the wider environment. Without this, regulatory frameworks will continue to rely on forms of evaluation that do not

[3] See DSIT's *Introduction to AI assurance* and the House of Commons Science, Innovation and Technology Committee's Governance of artificial intelligence (AI) Third Report.

capture the aspects of system behaviour that matter most. This requires investment in evaluation methodology – we call on the government to support research by statisticians and the broader mathematical sciences to develop novel methodological approaches to evaluating AI. We also recommend the development of a new government "colour book" for AI, providing practical guidance on the evaluation, deployment and use of AI across the public sector. Produced with significant input from statisticians and mathematical scientists, and maintained as a living document, it should evolve alongside the technology itself.

Recommendation 3: Ensure regulators can require evidence generation for high-risk AI systems

The case studies we discuss show that many of the most important behaviours of AI systems – eg, how outputs vary across users, inputs, and contexts – are not captured in existing documentation or logged data. Firms are often not incentivised to generate evidence that would reveal potential harms, and observed outputs alone are insufficient to assess how systems behave under alternative conditions. As a result, regulators may lack access to the information needed to identify discriminatory, anti-competitive, or otherwise harmful outcomes. Regulators should have the power to require firms to generate evidence about the performance of their AI models. In the case of AIs being used on health data, this might mean checking how an AI would respond to a re-identification attempt on an anonymised health dataset; in the case of AI tutors, this might mean assessing the reliability of a tutor's guidance and how well-matched it is to a student. This would enable oversight based on evidence of real-world system behaviour, rather than relying solely on pre-existing materials that do not capture the full range of risks.

Recommendation 4: Build statistical capability and embed statistical thinking within regulators

Many of the key challenges posed by AI – such as understanding reliability, assessing variation in performance across groups, and evaluating how systems behave as conditions change – are fundamentally statistical in nature. However, existing regulatory frameworks are not consistently equipped to address these issues. Regulators typically focus on processes and outcomes, but may not have the tools or expertise needed to interpret the behaviour of AI systems as statistical models. This capability is particularly important where risks are not fixed properties of systems: eg, where identifiability or reproducibility depends on evolving data, models and analytical techniques. Strengthening statistical capability within regulators is therefore essential. This includes access to appropriate expertise, methods for assessing uncertainty and robustness, and the ability to critically evaluate evidence on system performance. This may require establishing dedicated statistical functions within regulators, creating shared pools of specialist expertise across regulatory bodies, and developing common frameworks for evaluating AI systems that can be adapted to different sectors. Without this capability, regulators will struggle to set appropriate expectations, interpret evaluation results, or identify emerging risks in practice. Given the speed at which AI is moving, we regard this as an urgent need that should be addressed quickly.

Recommendation 5: Regulators should provide clear statistical guidance to organisations on the evaluation of AI

In many cases regulators do not directly regulate the use of AI within their sector: they regulate the outcomes of processes that use AI. The importance of AI models being evaluated according to sound statistical methodology – which emphasises their robustness, sensitivity to inputs, variation in performance across groups, and conditions under which outputs may be unreliable and not interpretable – is not fully appreciated by many of the organisations using AI. Regulators should help organisations to ask the right questions of developers, understand the assumptions underlying models, and identify the limitations of system outputs in practice. At a minimum, this guidance should empower organisations to

check that: 1) AI systems are evaluated in the contexts they're actually deployed in, not just against benchmark data; 2) policy or procedural decisions that guide an AI's behaviour are transparent and available for audit; and, 3) there is a defined method for testing whether a replacement or updated model matches the standard of the one it is replacing before it goes live. Without such guidance, there is a risk that firms are not equipped to assess whether the AI systems they use are consistent with regulatory expectations, particularly where those expectations depend on statistical behaviour that is not well understood.

2. AI in healthcare

One area of AI application where there are both significant opportunities and serious risks is the use of health data. The UK's health data assets are especially rich – spanning clinical records, longitudinal studies and genomics. AI models applied to this data can help identify patterns that improve diagnosis and treatment, and, when combined with wider administrative data, could also deepen understanding of environmental and social drivers of disease.

However, there are also risks. As the predictive power of AI increases, so too does the likelihood that individuals can be identified from seemingly anonymised data. The use of AI within scientific research also raises new challenges for reproducibility, and therefore for the reliability of evidence. These risks go to the heart of the regulatory challenge: if they are not addressed, the benefits of AI will be harder to realise and trust in its use will be undermined.

AI and de-anonymisation

A central challenge for regulators is the increasing difficulty of ensuring that health data can remain genuinely anonymised in the presence of modern AI systems. Maintaining public trust in this is essential if we are to maintain the consensus in favour of the use of anonymised data for health research. Traditional approaches to anonymisation rely on removing direct identifiers and limiting quasi-identifiers. This assumes that individuals cannot be identified from what remains – but this assumption is becoming increasingly fragile.

AI systems applied to rich datasets (such as those combining clinical, demographic and genomic information) can detect complex patterns that may effectively single out individuals, even where no obvious identifiers are present. Rare combinations of attributes, correlations across variables, or inferred characteristics could all enable an AI to identify an individual. This applies both to anonymised datasets that are shared with researchers and when AIs are used within trusted research environments: individuals' identity may be inferred through the modelling process. There are also risks associated with the outputs from research conducted in a trusted research environment: summary statistics, model parameters and synthetic data may appear safe in isolation but, when combined with other data sources, might enable re-identification.

The key change brought by powerful AIs is that identifiability becomes dependent on the interaction between data, the models used to analyse it, and the wider data environment. As AI models have improved along with data linkage and access, the question of whether an individual is identifiable from a dataset now needs to be treated as a context-sensitive risk calculation – considering how likely it is that an individual might be identified under plausible conditions and recognising that this might vary over time and contexts. Assessing the impact of an AI model on the risk of re-identification is a complex and fundamentally statistical question.

AI and reproducibility

A second challenge concerns the implications of AI for the reproducibility of scientific research. In the traditional model, reproducibility means that researchers can access the same data, and that applying the same methods yields the same results. This assumes analytical processes are transparent, well-defined and deterministic. That model needs to be reconsidered when advanced AIs are used as part of the research process.

This is because AI-based research depends on factors that are difficult to standardise or reproduce: their training data may be restricted or proprietary; their training process may include stochastic elements; and, results may depend on levels of computational resource (which not every researcher will have equal access to). This means that even where researchers share the models and code that they have used in their work, other researchers won't always be able to straightforwardly reproduce their results.

This issue arises because AI systems – especially the most powerful models -- cannot offer clear accounts of how their outputs are generated. They produce results based on statistical associations learned from data. This changes how we should understand reproducibility in AI-powered research: instead of research following clearly specified steps that can be repeated, results are generated by models whose behaviour depends on data, training processes and computational context. In this setting, reproducibility in the traditional sense (re-running a method and obtaining the same result) is no longer a realistic general standard. Results will vary as data, models and computational conditions change, and this variability is intrinsic to the way AI systems operate. There is a trade-off when using AI in this context: systems that are tightly controlled and fully reproducible will be more limited in scope; the more flexible and powerful approaches are capable of identifying complex or unexpected patterns but they are less stable and harder to replicate exactly. If the more powerful AIs are used in scientific research, we cannot expect strict reproducibility and instead must ask how stable and reliable AI is across plausible variations. This requires the development of explicit statistical measures of reproducibility (capturing sensitivity, robustness and uncertainty) which can be reported and assessed as part of the research process.

Implications for regulation

There are multiple regulators with responsibility for aspects of healthcare regulation – each has its own distinct mandate. The Information Commissioner's Office (ICO) oversees data protection, including anonymisation; the Health Research Authority (HRA) is responsible for research governance and ethics; the Medicines and Healthcare Regulatory Agency (MHRA) regulates AI systems where they function as medical devices; and NHS bodies set rules for data access and use. Each has responsibility for a part of the challenge outlined here, but none is responsible for assessing the statistical behaviour and reliability of AI systems in research.

The question of how likely it is that an individual can be identified from a dataset is becoming increasingly complicated: it depends on the dataset itself, the datasets it can be linked to and the types of AI tools that use the data. This means that there is a need to actively evaluate how the risk of reidentification changes as modelling techniques evolve and to test against realistic attempts at re-identification. Assessments of risk should become more empirical and dynamic to properly evaluate risk: this is an area where there is a crucial role for statisticians.

In relation to reproducibility, existing research governance emphasises transparency and good practice, but it is largely built around conventional understandings of what reproducibility means and how it is assessed. As AI methods become more central to research, this assumption becomes less tenable.

Research outputs using more powerful AIs are harder to reproduce exactly, but there is currently no established framework for evaluating reproducibility as a statistical property of AI-based research or for assessing the acceptability of any trade-offs. While regulators and funders encourage good practice, they do not typically require quantified evidence of stability, robustness or uncertainty as a condition of demonstrating reproducibility. Reproducibility is generally treated as something that follows from good research practice and transparency, rather than as a property of AI-based research that is actively evaluated. Our view is that there needs to be a shift towards assessing reproducibility as a property of AI-based research.

These issues suggest that the current framework is not well suited to systems whose behaviour is inherently statistical and context-dependent. The core gap is not simply a lack of oversight, but the absence of a consistent approach to evaluating and communicating uncertainty, variability and reliability. Addressing these challenges will require both stronger statistical capability within regulators and clearer standards for how AI systems used in research are evaluated and reported.

3. AI in education

The second area of AI application where there are both significant opportunities and emerging risks is in the context of education. There are a range of challenges in this domain – and they apply to schools, technical qualifications and higher education. Here, to give an idea of the challenges, we focus on schools and look at just two ways in which AI is being used in education. First, we look at the increasing use of AI tools by teachers to support tasks such as curriculum design, lesson planning, generating teaching materials, designing assessments and marking. Second, we look at AI tutors where there is growing interest in their use to provide more direct, adaptive support to students. These applications have the potential to improve efficiency and consistency in teaching, and to reduce administrative burdens, allowing teachers to focus more time on direct engagement with students. AI tutors potentially enable more students to benefit from tailored one-to-one support.

However, these opportunities are accompanied by risks. As the use of AI becomes more embedded in teaching, assessment and learning processes, the quality, reliability and appropriateness of the outputs these systems produce becomes increasingly important. Unlike traditional resources, which are often developed, reviewed and refined over time, AI-generated content is produced dynamically and may vary across uses. The behaviour of these systems may also change as they adapt to different inputs or contexts. This means that an AI tutoring tool cannot only be evaluated when it is launched – as the way in which it develops its response through interaction with a student could affect its accuracy, fairness and educational quality.

Though the issues that we discuss here are specific to education, similar issues arise across the public sector where AIs influence outcomes for individuals: the lessons are transferable to child protection, policing, parole, welfare decisions and so on.

AI and teaching practice

One key regulatory challenge is to ensure that the AI systems used by teachers maintain educational quality. AI tools generate content that is fluent, plausible and internally consistent – but, as we set out in *AI is Statistics*, the content is not necessarily accurate, unbiased or appropriate for every context in which it might be used. When such outputs are incorporated into lesson plans, teaching materials or assessment processes, errors or distortions may be embedded directly into classroom practice.

AI-generated examples may be factually incorrect or misleading, but difficult to identify as such. Content may reflect patterns in training data that are not representative of the student population, resulting in examples or explanations that resonate with some groups of students more than others. When used in marking or feedback, AI tools may introduce systematic inconsistencies or biases, particularly where their behaviour varies across different types of input.

These would not be isolated errors; they would reflect underlying statistical properties of the systems themselves. The outputs produced by AI models depend on their training data, model structure and the prompts they receive, and may vary in ways that are not transparent to users. In particular, these systems do not typically communicate the degree of uncertainty associated with their outputs, or signal when they are operating outside the contexts for which they are most reliable. As a result, teachers may rely on outputs without a clear understanding of their limitations. If these are to be used by teachers, the AI models should come with some indication of their reliability and appropriateness to different circumstances and it must be possible to assess them in practice.

AI tutors

A second challenge arises from the use of AI systems as tutors or learning assistants. This is especially important to highlight, as government has indicated that these tools are going to be rolled out with their backing. AI tutors go beyond generating static content: they interact dynamically with students and adapt responses based on their inputs. This means that they make ongoing assessments of students' levels of understanding and how best to support their learning. But these assessments will have degrees of uncertainty – and these should be visible to users. Our understanding of the currently available AI tutors is that they do not typically communicate how confident they are in their responses, how reliable their guidance is, or how well it is matched to the needs of the individual student. These issues are exacerbated by the manner in which current AI tutors communicate with students – for example, they might classify answers as incorrect if students skipped a step or ask follow-up questions that are too open-ended to effectively support independent learning.

There is a distinct statistical challenge here. The AI model infers a student's knowledge from limited information, and uses that to guide subsequent teaching. Errors in this process, whether through incorrect explanations, misjudgement of ability or inappropriate sequencing of material, risk compounding over time as the system continues to adapt – risking students developing misunderstandings or disengaging with the material. At the same time, there is limited evidence on how these systems perform across different groups of students or contexts. As a result, AI tutors may give the appearance of personalised and reliable instruction, while in practice operating with uncertainty and variability that is not visible or understood.

Implications for regulation

Our view is that AI tools should be evaluated to ensure that they meet standards of reliability and fairness before they are used in schools. However, it is not clear that, in England, there is a regulator with responsibility to do this. Ofsted is responsible for assessing the quality of education provided by schools while Ofqual's role focuses on assessment and qualifications. Neither regulator has an explicit mandate to assess the reliability of AI systems as they are used within teaching and assessment.

In practice, the use of AI in teaching is largely left to the judgement of schools and individual teachers. While this reflects an appropriate degree of professional autonomy, it also means that there is no consistent approach to ensuring that AI-generated content is sufficiently accurate, appropriate or fair for

its intended purpose. Guidance to teachers on how to evaluate AI tools is limited, and there is little transparency about how such tools have been assessed or what their limitations may be.

This creates a gap. AI is being incorporated into core educational processes (lesson design, content delivery and assessment) but there is no consistent, sector-wide approach to evaluating whether the outputs these systems produce are reliable enough for those uses. Responsibility for judging quality therefore rests with individual teachers, who do not have the information or tools needed to do so consistently.

The lack of regulation is especially acute in the case of AI tutors, where systems interact directly with students and influence learning pathways over time. The challenge is not just about the accuracy of individual outputs – it is about how we evaluate the behaviour of the system as it adapts and makes implicit judgements about students 'understanding. At present, there is no clear framework for evaluating these systems in practice, or for ensuring that their behaviour is sufficiently reliable for use in education.

It is worth noting that we have not discussed the issue of AI being used by students to complete tasks – as that is (largely) not an issue for regulators. Where there is a potential regulatory angle is if education providers start to use AI-powered tools to determine whether students have used AI in their work. It is worth noting that the EU classifies AIs that perform this role as high-risk and regulate them accordingly. It is right that this is a high-impact use of AI and we see this as another area where statistical evaluation is key and there is a role for regulatory oversight.

The problem here is primarily that there is a gap in responsibility. AI systems are being used in teaching and assessment without any consistent requirement to demonstrate that their outputs are sufficiently accurate, appropriate or fair for those purposes. There is a job for some regulator to set clear expectations that AI tools used in schools are subject to appropriate evaluation before and during use. This should include requiring evidence about the reliability of outputs, the potential for systematic bias, and the contexts in which the tools can be used safely. It should also ensure that teachers are provided with clear information about these limitations, so that professional judgement is supported rather than undermined. At present, no part of the system performs this role, and addressing this will require an explicit decision about where that responsibility should sit.

As indicated at the start of this section, there are similar issues arising wherever AI is used in the public sector. Particularly the challenge around regulatory gaps. This is why our key recommendation is that DSIT should take clearer responsibility for identifying where statistical risks from AI are not being covered and for ensuring that the most appropriate regulator is required (and equipped) to address gaps.

In our second recommendation we have highlighted the need to invest in evaluation of AI models – we see education as an area where this is especially urgent. We urge the government to engage with statisticians and other mathematical scientists to develop robust approaches for evaluating AIs that are designed for use in education – this needs to include engaging with education researchers, teachers and students. We urge the government to fund this work to develop new evaluation methodology and to commit to a new "colour book" to provide guidance across the public sector.

4. AI in finance

The final area of AI application that we are highlighting is the provision of financial advice and support for financial decision-making – though, while we specifically discuss financial advice here, the regulatory lessons transfer across to any sector where AI provides advice. AI tools are increasingly used to help

individuals manage budgets, choose financial products and make decisions about saving, borrowing and investing. These applications have the potential to improve access to financial guidance, particularly for individuals who may not otherwise receive tailored advice, and to provide more responsive, personalised support. In the UK, this includes the development of "targeted support", which sits between personalised financial advice and general guidance and allows firms to provide recommendations tailored to groups of people with similar characteristics, rather than to individuals.

However, as in the case of education and health data, there are significant risks. Financial decisions are inherently uncertain, and the consequences of poor advice are potentially life changing. As AI becomes more widely used to provide or support financial guidance, the reliability, robustness and appropriateness of its outputs becomes critical. While such systems may appear consistent and authoritative, there is a risk that they are deployed without a clear understanding of how well they perform in practice.

AI and the evaluation of financial advice

A central challenge arises from how these systems are evaluated. In many cases – as detailed in AI is Statistics – AI models are assessed using benchmark datasets or controlled environments, which may not reflect the contexts in which they are used in practice. This means that evaluation can give a misleading picture of how systems will perform under real-world conditions. This creates a set of underlying statistical issues:

- Evaluation data may differ systematically from real-world conditions, leading to an unreliable estimation of performance in real-world conditions.
- Systems may be sensitive to relatively small changes in inputs, producing different recommendations without users being aware of how a small change in their input may have produced different advice.
- Performance may also vary across different groups of users, particularly where training data does not reflect the diversity of financial circumstances.
- Even when accurate and reliable, an AI's outputs and predictions may not be explainable and financially interpretable, preventing control and monitoring of its consequences.
- There can be a mismatch between what is measured and what matters: systems can perform well on benchmark tasks while still providing advice that is unreliable or inappropriate in real-world settings.

Another major challenge is that evaluation is often conducted as a one-off exercise prior to deployment. In practice, financial conditions, user behaviour and underlying data all change over time, meaning that the performance of systems may degrade or vary in ways that are not captured by initial testing. If the evaluation of AIs producing financial advice is conducted in this way, it would mean that there is no consistent way of assessing whether an AI is fit for purpose in practice.

These issues have another dimension in the context of the emerging forms of "targeted" financial support. Unlike traditional financial advice, which is tailored to an individual, or general guidance, which must be broadly applicable, these systems provide recommendations to groups of users defined by shared characteristics. In doing so, they rely on statistical assumptions about how "people like you" behave or what outcomes are likely. The quality of the advice provided therefore depends not only on

overall performance, but on how well these group-level inferences reflect real-world variation across users and contexts.

These challenges have important implications for financial inclusion. The potential of AI is to produce tailored advice for individuals or groups with similar characteristics – potentially enabling people who would not usually access good financial advice to access it. But, if these systems are evaluated using unrepresentative data or are not assessed for how performance varies across different groups, there is a risk that they systematically work less well for those with more complex or less typical financial circumstances. In this setting, weaknesses in evaluation are not just technical issues, but can translate directly into uneven outcomes, with some groups receiving advice that is less accurate or appropriate despite increased access to support.

Implications for regulation

These challenges take on a specific form in financial services. The Financial Conduct Authority (FCA) strictly regulates advice given by firms to individuals and is developing an approach for firms offering targeted support to groups of consumers. However, regulation focuses primarily on the advice provided, rather than the tools used to generate it. This means that firms remain responsible for outcomes, but risks may still arise where AI systems are introduced into workflows without being properly evaluated or monitored over time. While it is neither feasible nor desirable for the FCA to regulate every AI system used by firms, there is a strong case for regulators to provide clearer guidance on what organisations should expect from AI tools and how they should assess them before use.

A distinct challenge arises in the context of targeted support. These systems operate by grouping individuals and applying statistical assumptions about how people with similar characteristics are likely to behave or what outcomes are appropriate. The quality of advice therefore depends on how well these group-level assumptions reflect real-world variation. However, such systems do not typically make transparent how groupings are constructed, how sensitive recommendations are to variation within groups, or the circumstances in which the advice may be less reliable. This creates a gap between how the systems operate and how they are currently assessed. The FCA's guidance for targeted support should emphasise the importance of firms being transparent about these statistical aspects of their advice.

For the FCA (and other regulators in a similar position) to regulate AI effectively, they will need to engage more directly in future with questions of statistical methodology. This applies both to evaluating systems within their remit and to supporting the organisations they regulate in assessing the tools they use. In practice, this will require stronger statistical capability across the regulatory landscape, alongside clearer guidance on how AI systems should be evaluated and deployed in practice.

5. Conclusion

Taken together, these three case studies illustrate a broader challenge: effective AI regulation depends on translating highly specialised domain knowledge into robust statistical methods for evaluating system performance. They show that current frameworks do not consistently engage with the statistical properties of these systems, or provide a clear basis for assessing how they behave in practice. Addressing this requires a more coordinated approach across government, with clear responsibility for identifying where uses of AI fall between regulatory remits, stronger expectations that evaluation reflects statistical characteristics such as uncertainty and variation, investment in statistical capability within regulators, and clearer guidance to the organisations increasingly using AI on how and when such

evaluation is necessary. Without this, there is a risk that AI systems are deployed at scale without demonstrating that they are reliable, appropriate or fair in practice – this in turn risks harming public confidence and limiting our ability to reap the benefits of the technology.

From past to present...

The image of the wheatsheaf first appeared in our original seal. Being the end product of the harvesting and bundling of wheat, it was a pictorial way of expressing the gathering and analysis of data: the foundations of statistical work.

It also implied that statistical practice comprises more than the collection of data: it consists of active interpretation and application as well (threshed for others, if the rural analogy is sustained). Rigorous data gathering is still at the heart of modern statistics, but as statisticians we also interpret, explain and present the data we collect.